Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

On the Analysis of COVID19 - Novel Corona Viral Disease Pandemic Spread Data Using Machine Learning Techniques

Shreyas Setlur Arun
Department of Computer Science and Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
shreyas.s.a@live.com

Ganesh Neelakanta Iyer
Department of Computer Science and Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
ganesh@ganeshniyer.com

Abstract—**Coronaviruses are a group of viruses that cause various diseases in mammals and birds. In humans, they cause a range of respiratory disorders. This paper presents the analysis of the transmission of COVID19 disease and predicts the scale of the pandemic, the recovery rate as well as the fatality rate. We have used some of the well-known machine learning techniques as well as mathematical modeling techniques such as Rough Set-Support Vector Machine (RS-SVM), Bayesian Ridge and Polynomial Regression, SIR model, and RNN.**

Keywords—COVID19, Support Vector Machine, Bayesian Ridge, Polynomial Regression, SIR Model, Coronavirus, Machine Learning, Recurrent Neural Networks, Long Short-Term Memory

I. INTRODUCTION

Viruses are microscopic organisms that replicate only inside the living cells of an organism. Coronaviruses are one of the most common families of viruses that cause various respiratory diseases in living organisms.

Forecasting the spread of coronavirus is one of the challenges in recent times. Forecasting the pandemic with high accuracy will help different countries prepare a plan to fight a war against the virus spread. Machine learning techniques are extensively used for modeling real-world problems [13], [14]. Specifically, applications of machine learning techniques to predict diseases have been used extensively in recent times [4],[5],[8].

Some of the traditional methods used in forecasting of an epidemic include time series modeling as well as regression modeling. We have chosen some of the traditional as well as modern machine learning techniques. Support Vector Machine technique has been traditionally used for predictions while we have compared some of these traditional techniques with a modern technique namely Recurrent Neural Network (RNN). As there is always a great degree of uncertainty in the prediction of pandemics we have also calculated the Mean Square Error (MSE) and Mean Absolute Error (MAE) using the Bayesian ridge technique. The use of polynomial regression method was to find the best relationships between the variables present in the data. We have attempted to solve the problem of forecasting the pandemic of COVID19 using the machine learning methods stated above.

A reminder of the paper is organized as follows. Section II briefly explains the literature review, section III describes the machine learning methods used, section IV describes the results in detail, and finally section V concludes the paper.

II. LITERATURE SURVEY

As per our knowledge mortality rate of heart, the brain, as well as kidney diseases, has been done frequently whereas the spread of
infectious diseases is not too popular. In this section, we focus on giving information about similar initiatives taken.

Andrea et.al [1] propose a mathematical expression to predict the spread of H1N1 influenza pandemic. Transmissibility of the pathogen, the transportation network and mobility features, the demographic profile, and the mixing pattern of the virus are the deciding factors. The results are classified according to the age group and travel history of the people.

Stephen et.al [6] propose a mathematical framework using actual census, land-use, and population-mobility data to model disease outbreaks. The result is published in the form of a bipartite graph mapping the number of locations visited, the number of visitors, and the number of contacts.

Randi et.al [7] propose a graphical structure for the spread of infectious diseases in a primate social network based on Eigenvector centralization index. The graphical structure is based on the interaction between an individual and a group. It achieves an accuracy of 72.6%.

Anjan et.al [9] propose work on predicting the mortality of heart diseases using classification by Naive Bayesian classifier. They have also employed advanced encryption techniques for securing the personal data of patients. The

reported accuracy is 84.07% for classification and 98.2% for the security of data.
Olivera et.al [10] propose a model based on Bayesian Monte Carlo probabilistic theory to predict the spread of infectious diseases. The major disadvantage of this method is that it gets affected by the selection of prior as well as posterior probabilities.

Caroline et.al [11] propose a mathematical model for the prediction of infectious diseases. They have focused on the spread of influenza transmission between birds and humans. The accuracy of the mathematical model is measured based on the pragmatic form of validation.

III. METHODOLOGY

In this section, we explain our model in detail. Initially, we brief the data set that we have obtained from Johns Hopkins Center for Systems Science and Engineering and the pre-processing techniques explored. Then we explain the machine learning technique for predicting the spread of the pandemic.

A. Dataset preparation and data pre-processing

The data set used in this project was the real-time data from Johns Hopkins Center for Systems Science and Engineering [12]. It is the real time data of the patients. The data consists of the regions of the state, the country, the latitude, the symptoms, the longitude, and the number of cases for each day starting from 22/01/2020. We have considered the updated data set for the project until 07/05/2020. The steps followed for preprocessing of the data is as follows:

- The data was cleaned and the null values were replaced by averaging the column.
- The data was transformed using the StandardScaler object in Python to achieve a Gaussian distribution for predicting the spread of the pandemic.
- The data were normalized using the logarithmic scale for removing the outliers.

B. Mathematical modeling using SIR Model

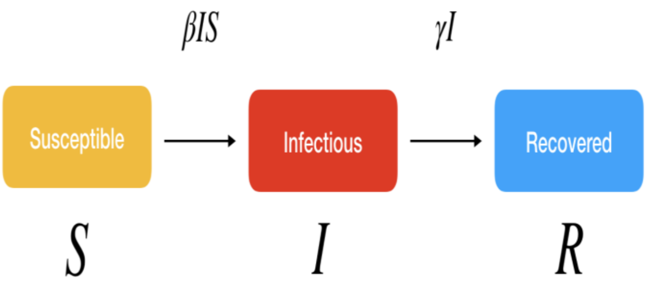

Fig.1 Representation of SIR Model using a block diagram

SIR Model [3] is a theoretical mathematical model used to predict infectious diseases. It is based on the differential equation modeling. The model is diagrammatically represented in Fig.1 and the defined as follows:

- Goal: The model aims to find the relation between the dependent and the independent variable.
- The independent variable is considered to be time 't'.
- One of the groups considers individuals whereas the other group considers a fraction of the population.
- We have considered the group consisting of a fraction of the population for better accuracy.
- The model is defined as follows:
  a) Susceptible (S): The individual hasn't contracted the disease, but she can be infected due to transmission from infected people.
  b) Infected (I): The individual has contracted the disease.
  c) Recovered/Deceased (R): The disease may lead to one of two destinies: either the person survives, hence developing immunity to the disease or the person is deceased.
- β is a parameter which gives the rate of transmission of the disease from one person to another. It is determined by the chance of contact and the probability of disease transmission.
- γ is a parameter which expresses the rate of recovery in a specific period.
- Once the people are healed, they get immunity. The people who die by natural death are not considered.
- D is assumed to be the number of days taken to recover from the infection. It is derived from γ.
- R_0 is the basic reproduction number of the disease. It gives the average number of people affected by another person. It is used to estimate the Herd Immune Threshold (HID).

$$ds/dt = -\beta IS \tag{1}$$

$$di/dt = \beta IS - \gamma I \tag{2}$$

$$dr/dt = \gamma I \tag{3}$$

$$D = 1/\gamma \tag{4}$$

Balanced state of the disease is obtained by multiplication of the basic reproduction number by the percentage of non-immune people and is equal to 1.

$$R_0 = \beta/\gamma \tag{5}$$

Let the number of immune people be p. Then the stable state of the pandemic can be represented as:

$$R_0(1 - p) = 1 \rightarrow 1 - p = 1/R_0 \rightarrow p_c = 1 - (1/R_0) \tag{6}$$

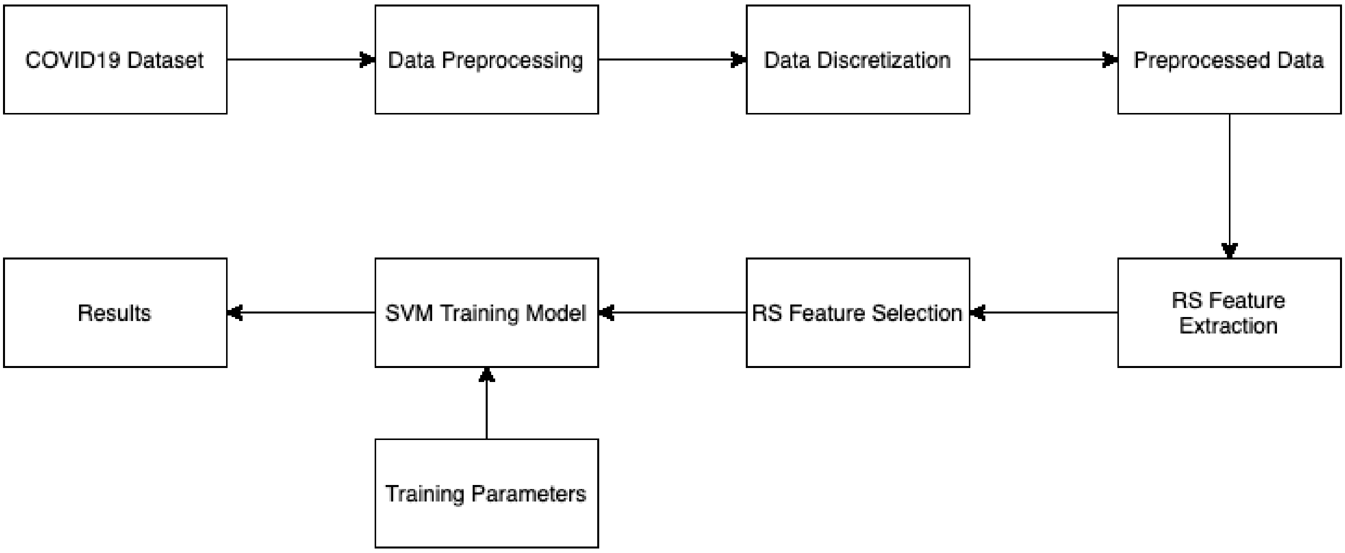

Fig. 2. Architecture of RS Based SVM System

C. Prediction using RS-SVM

SVM is a classical machine learning algorithm used for making predictions. We have used the Rough Set (RS) Attribute Reduction SVM. The RS is used as an anterior preprocessor of SVM to cut down the complexity of SVM. This method increases the accuracy of the prediction. The major idea of the RS based SVM algorithm is to reduce redundant and irrelevant information. Fig.2 represents the architecture of the RS based SVM system used in the experiment. It follows a heuristic approach and is classified under greedy algorithms. Algorithm 1 [3] was used in the experiment.

Algorithm 1 A Greedy Attribute Reduction Algorithm

```
Input: A decision system DS = < U, R = C ∪ E , V, f>
Output: A reduct of DS, denoted as Redu
P = POSc(E)
Put C into array c, put Φ into array Redu
K = 0
while pos <> k do
    j = 1
    for i ← 1 to |c| do
        if $POS_{Redu \cup C[i]}$ [E]
            K = $POS_{Redu \cup C[i]}$ [E]
            J = i
    Redu = Redu ∪ c[j]
    Delete the element c[j] from array c
    for i ← 1 to |Redu| do
        if $POS_{Redu - Redu[i]}$ (E) = P then
            Delete the element Redu[i] from array Redu
```

This algorithm works as follows:

- The reduction set is assumed to be empty.
- Criteria of classification are set according to the experiment.
- The attribute having the best classification ability is chosen.
- The experiment is repeated iteratively until a specified reduction set of attributes is obtained.

D. Prediction using Polynomial Regression

Polynomial regression is a quadratic form of linear regression [2]. The advantage of using polynomial regression is to overcome the dependency between variables which may not be linear [17].

Fig.3 depicts the architecture of the polynomial regression-based system used in this experiment. We have used the 5th-degree polynomial regression. The decision to implement 5th-degree polynomial regression was taken as the curve was able to fit the data precisely. The tradeoff between bias and variance was minimum in the 5th-degree polynomial regression prediction. Fig.3 depicts the architecture of the polynomial regression-based system used in the experiment.

E. Prediction using RNN

Recurrent Neural Networks (RNN) form a class of neural network algorithms having an internal memory in each hidden layer[16]. It stores information that has been previously calculated. RNN is recurrent as they use the same parameters for each batch of inputs in each hidden layer. The inputs are processed based on the weights and biases for each hidden layer. Once the output is produced, it is copied and sent back into the recurrent network. Long Short-Term Memory (LSTM)

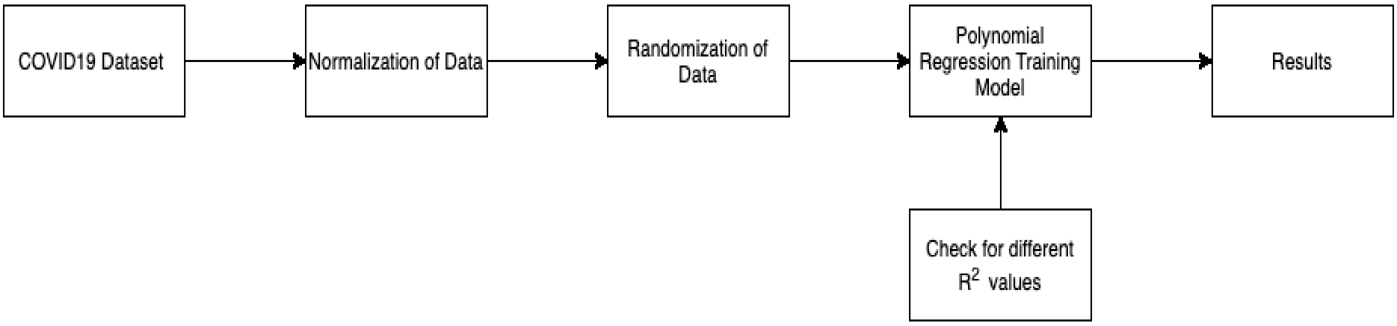
COVID19 Dataset
Normalization of Data
Randomization of Data
Polynomial Regression Training Model
Results
Check for different R^2 values

Fig. 3. Architecture of Polynomial Regression Based System

networks use an enhanced version of recurrent neural networks, which makes it easier to remember past data in memory [15]. We have used the LSTM technique to predict the rate of fatality of the pandemic. The major advantage of using LSTM is that it is an appropriate technique to classify, process, and predict
time series models having different lag durations. The model is trained using backpropagation mechanism.

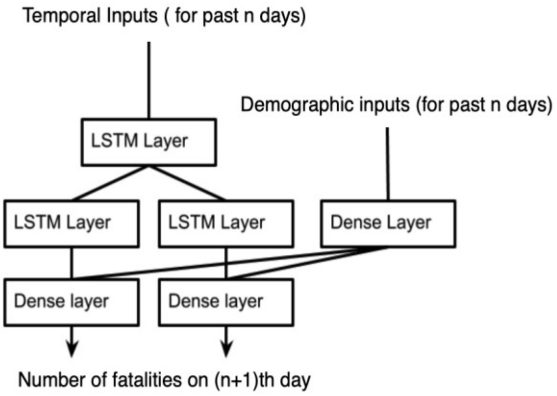
Temporal Inputs (for past n days)
Demographic inputs (for past n days)
LSTM Layer
LSTM Layer
LSTM Layer
Dense Layer
Dense layer
Dense layer
Number of fatalities on (n+1)th day

Fig. 4. RNN Architecture

Fig.4 details the architecture of the RNN model used in the experiment. 5 layers of temporal inputs, as well as 5 layers of demographic inputs, were considered for the experiment. 250 epochs were considered for greater accuracy of the prediction. The dataset was divided into a training set as well as a testing set.

F. Computation of MSE and MAE

The computation of Mean Square Error and Mean Absolute Error was very crucial in our experiment. The prediction of the spread of the pandemic was based on the Bayesian estimators. The Bayesian technique involves the assumption of the prior distribution which helps in minimizing the loss of the posterior expected value. An estimator is considered over the probability distribution and this acts as the risk function. The estimator is called the Bayes estimator if there is a reduction of Bayesian risk among all the estimators in the prediction. The loss function is assumed to be a real-valued function L(Θ, a). 'a' represents
the estimate whereas 'Θ' represents the true value of the parameter. The biggest advantage of using Bayesian estimator-
based technique was that it can be improved as newer data arrives in the database.

IV. Results and Discussion

A. SIR Model Results

The SIR model was used to estimate the parameters to calculate the rate of spread of the pandemic as well as the recovery in 4 countries namely Japan, South Korea, Iran, and Italy. Table I represents the results of the model and is shown below:

TABLE I. SIR Model Results Table

Country	β	γ	R_0
Japan	0.00002856	0.29819303	0.00009578
South Korea	0.00001297	0.00000001	1297.49430758
Italy	0.00001582	0.00000001	1581.92377423
Iran	0.00006294	0.3999237	0.00015735

RS based SVM technique was used to predict the rise in the number of cases of COVID19 as well as the comparison between the rate of recovery as well as the mortality of the disease.

- *Prediction of the spread of COVID19 cases:* RS based SVM technique was very accurate in the prediction of the spread of the COVID19 pandemic. The exponential rise in the number of infected cases was accurately predicted. Fig.5 details the results of the prediction and the model was able to achieve an accuracy of 85%.

- *Comparison between confirmed, recovered and mortality rates of COVID19 in India:* Fig.6 represents the prediction for the growth rates of the confirmed cases, recovery rate as well the death rate for the pandemic were calculated using the RS based

SVM technique using Indian data. We were able to achieve an accuracy of 86%.

The analysis that we will be able to conclude from the above two analyses is that the pandemic is highly infectious and deadly compared to the recovery rate of patients.

B. Polynomial Regression Results:

Polynomial regression was used to predict the rate of mortality as well as the rate of recovery. The linear regression technique was used to fit the data but was under fitted. The polynomial curve of the 5th degree was the perfect fit for the data.

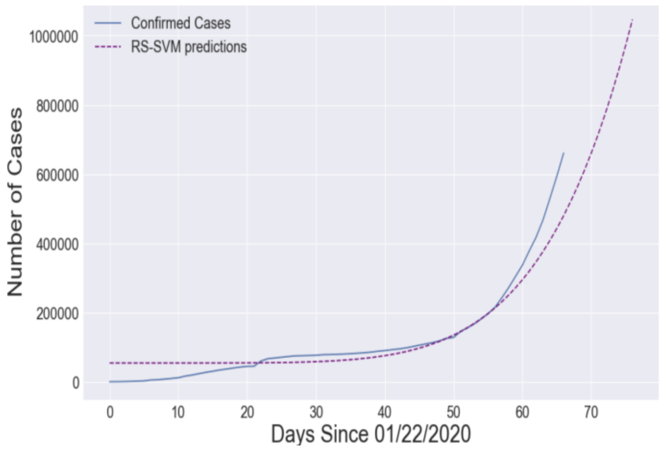

Fig. 5. Prediction of the spread of COVID19 cases

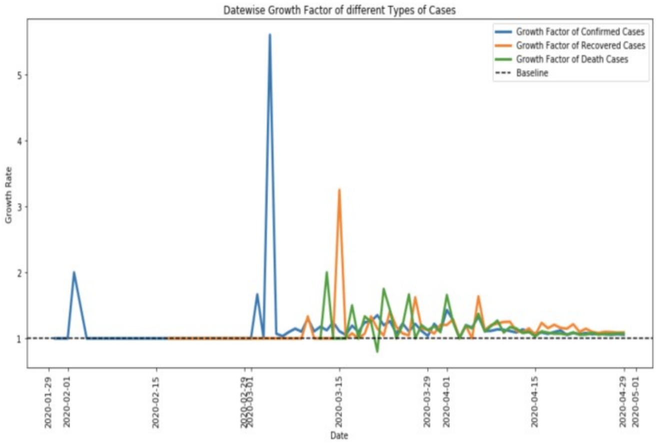

Fig. 6. Growth factor for COVID19 cases

- *Rate of Mortality:* Polynomial Regression was used to predict the rate of mortality for the 15 countries that have the highest cases of mortality. Fig.7 details the results of the prediction of the rate of mortality and the accuracy achieved in this case was 85.14%.

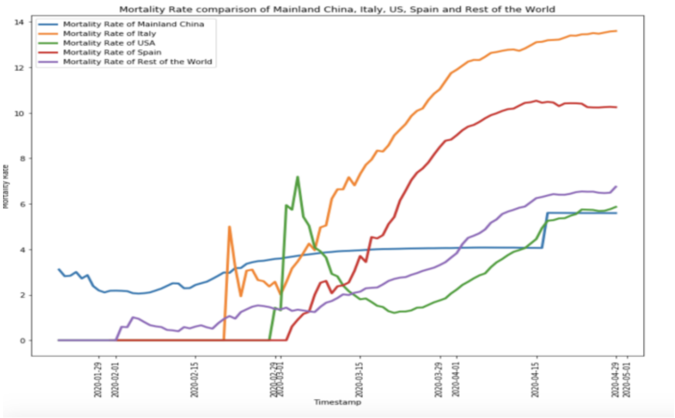

Fig. 7. Rate of Mortality

- *Rate of Recovery:* The rate of recovery of patients is one of the most important parameters for judging a pandemic. Fig.8 represents the recovery rate of the pandemic. The logarithmic values were considered to avoid outliers affecting the analysis. We were able to achieve an accuracy of 87.23% using polynomial regression.

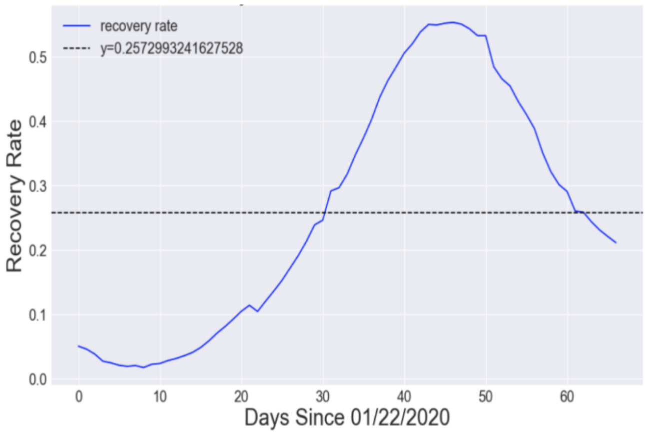

Fig. 8. Rate of Recovery

C. Recurrent Neural Network Result

The rate of recovery, the spread of the infection as well as the rate of mortality was calculated for each country using the RNN model. The overall accuracy of the model was around 89%. The usage of the LSTM technique helped in reducing the error rate due to the epochs. Fig 9. diagrammatically represents the loss function of the RNN used in the experiment.

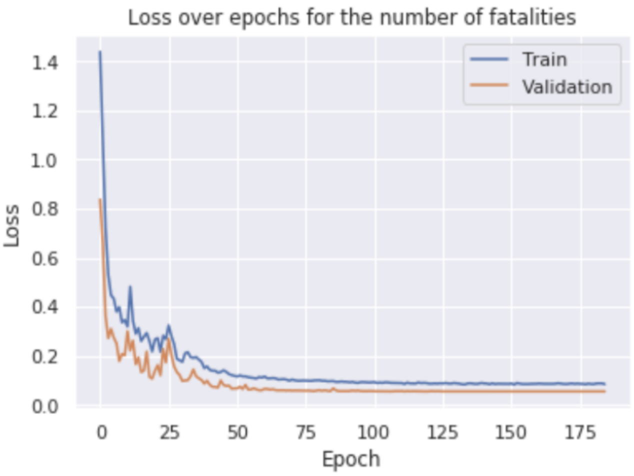

Fig. 9. Loss over epoch for the number of fatalities

D. Analysis of the spread of the pandemic

The COVID19 pandemic is a very fast-spreading infectious disease. The logarithmic scale was considered to overcome the outliers in the data. Table II details the confusion matrix for the machine learning techniques that have been used in the experiment.

Table II. CONFUSION MATRIX

ML Algorithm	Precision	Recall	F1 Score
RS-SVM	0.78	0.43	0.57
Polynomial Regression	0.69	0.57	0.61
RNN	0.81	0.41	0.55

After the computation of the results using the three machine learning algorithms, we were able to conclude that the RNN algorithm provided the best results. RNN was able to predict the results with greater accuracy and precision. Polynomial regression was affected by outliers in the data whereas RS-SVM was affected by overfitting and underfitting of the data.

V. CONCLUSION AND FUTURE REMARKS

In this paper, the analysis was performed based on the data set provided by the Johns Hopkins Corona Virus Resource Center. We had deployed some of the most popular machine learning algorithms and were able to achieve considerably good results. The proposed RNN model was the most efficient among the machine learning models deployed in the experiment.

The results are useful in predicting the spread of any epidemics or pandemics for any country or the whole world and containing it. The challenge in the analysis of this data set is that it is growing by the day and the number of cases is increasing exponentially. Several future works include the deployment of a combination of neural networks with traditional machine learning techniques.

REFERENCES

[1] Andrea Apolloni, Chiara Poletto, and Vittoria Colizza. “Age-specific contacts and travel patterns in the spatial spread of 2009 h1n1 influenza Pandemic”, BMC infectious diseases, 13(1):176, 2013.

[2] WA Bergerud. “Introduction to regression models: with worked forestry Examples”. biom. info. hand. 7. Res. Br., BC Min. For., Victoria, BC Work. Pap, 26:1996, 1996.

[3] Pratchaya Chanprasopchai, I Ming Tang, and Puntani Pongsumpun, “ Sir model for dengue disease with effect of dengue vaccination.”, Computational and mathematical methods in medicine, 2018, 2018

[4] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang., “Disease prediction by machine learning over big data from healthcare communities”, Ieee Access, 5:8869–8879, 2017.

[5] Dhiraj Dahiwade, Gajanan Patle, and Ektaa Meshram, “Designing disease prediction model using machine learning approach”, In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pages 1211–1215. IEEE, 2019.

[6] Stephen Eubank, Hasan Guclu, VS Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. Nature, 429(6988):180–184, 2004.

[7] Randi H Griffin and Charles L Nunn, “Community structure and the spread of infectious disease in primate social networks”, Evolutionary Ecology, 26(4):779–800, 2012.

[8] Pahulpreet Singh Kohli and Shriya Arora, “Application of machine learning in disease prediction”, In 2018 4th International Conference on Computing Communication and Automation (ICCCA), pages 1–4. IEEE, 2018

[9] Anjan Nikhil Repaka, Sai Deepak Ravikanti, and Ramya G Franklin. “Design and implementing heart disease prediction using naives bayesian”, In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pages 292–297. IEEE, 2019.

[10] Olivera Stojanovi´c, Johannes Leugering, Gordon Pipa, St´ephane Ghozzi, and Alexander Ullrich. A bayesian monte carlo approach for predicting the spread of infectious diseases. PloS one, 14(12), 2019.

[11] Caroline E Walters, Margaux MI Mesl´e, and Ian M Hall. Modelling the global spread of diseases: A review of current practice and capability. Epidemics, 25:1–8, 2018.

[12] https://github.com/CSSEGISandData/COVID-19, Last accessed: 7st May 2020

[13] Bhavanam Lakshmi Tulasi, Dr. Ganesh Neelakanta Iyer, "On the Classification of Kathakali Hand Gestures Using Support Vector Machines and Convolution Neural Networks", in International Conference on Artificial Intelligence and Signal Processing (AISP) 2020, VIT, Amaravathi, India, 2020

[14] Sanjay Kumar K. K R., Goutham Subramani, Rishinathh K. S., and Dr. Ganesh Neelakanta Iyer, “On multi-class currency classification using mind”, in Springer International Conference on Advances in Distributed Computing and Machine Learning(ICADCML-2020), VIT, Vellore, India, VIT, Vellore, India, 202

[15] Vijayakumar T, “Comparitive Study Of A Capsule Nueral Network In Various Applications”, Journal of Artificial Intelligence, 1(01), 19-27, 2019

[16] Raj, J. S., & Ananthi, J. V,”Recureent Nueral Networks and Non Linear Prediction In Support Vector Machines”, Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40, 2019

[17] S.A. Shreyas, S Abishek , N. Radhika, “Performance Analysis of Linear Grid Stability Using Classifiers and Advanced Ensemble Techniques”, Journal of Advance Research in Dynamical & Control Systems, Vol. 11, 03-Special Issue, 2019